Research Article

Harnessing the Power of Corpus Linguistics in Language Education: A Student-Centered Approach

Katerina Florou[1]*

1 National and Kapodistrian University of Athens, Athens, Greece

*** Corresponding author:** Katerina Florou, Department of Italian Language and Literature, Faculty of Philosophy, National and Kapodistrian University of Athens, Greece. Tel: +302107277491; Email: kathyflorou@ill.uoa.gr

A B S T R A C T

Introduction: Corpus linguistics (CL) has emerged as a valuable approach in language education, yet skepticism persists among future foreign language teachers regarding its necessity and effectiveness. This study investigates whether hands-on experience with corpora and natural language processing (NLP) tools can shift pre-service teachers' perceptions and enhance their appreciation for corpus-based methods in language teaching.

Methodology: A structured teacher-training course was implemented, during which participants created their own corpora and utilized NLP tools to develop teaching materials. Pre- and post-course questionnaires were administered to assess changes in attitudes towards CL. The chi-square tests were used to analyze the significance of the collected data.

Results: Findings indicated a significant positive shift in perceptions. Engagement with CL tools led to increased appreciation for data-driven methodologies, with participants expressing a greater likelihood of incorporating these tools into their future teaching practices. The chi-square analysis confirmed the statistical significance of these changes.

Conclusion: Practical engagement with CL and related technologies can effectively address initial skepticism among teacher trainees. These results advocate for the inclusion of CL components in teacher training curricula to promote innovative, data-driven language teaching practices and bridge the gap between skepticism and effective application.

1. Introduction

In the context of corpus linguistics, when explaining to future language teachers how to apply corpora in a data driven learning (DDL) in foreign language teaching, there are some doubts about the necessity and the effectiveness of the task (Boulton & Tyne, 2014). These doubts often align with longstanding critiques of corpus-based approaches. For instance, Flowerdew (2005) contends that corpus linguists frequently overlook the socio-cultural dimensions of language, as they work predominantly with decontextualized data.

A substantial body of research underscores concerns regarding the practical integration of corpus-drivenK strategies in classroom settings. For instance, Boulton (2008, p. 3) asserts that "learners

and teachers simply aren't convinced" about the use of corpora in foreign language teaching. He further notes that students often struggle with mastering collocations, a core element of corpus-based learning (Boulton, 2009). Similarly, Yoon and Hirvela (2004) report findings in which a large percentage of participants experienced difficulties when studying keywords within collocational frameworks.

Teacher perceptions further complicate implementation. As Karlsen (2021) indicates, educators' doubts about students' linguistic proficiency and digital literacy can deter them from adopting corpus-based methodologies. This resistance is compounded by teachers' previous exposure to corpora, which is frequently confined to

theoretical or research-oriented contexts, thereby reinforcing a perception of corpora as inaccessible or irrelevant to everyday pedagogy (Farr, 2008). Collectively, these interconnected factors, skepticism among both students and teachers, contribute to the broader reluctance to embrace corpus-based approaches within English as an Additional Language (EAL) instruction.

On the other hand, a substantial body of literature advocates for the integration of Corpus Linguistics (CL) in foreign language education, particularly within the context of English as a Foreign Language (EFL). One of the early contributions in this domain is Aston's (2022) chapter, which introduces the concept of students as corpus creators. Building on this premise, Gavioli (2009) emphasizes the role of corpora in fostering learner autonomy. Further practical guidance is provided by Reppen (2010), who outlines effective strategies for employing corpora in classroom instruction, and by De Cock (2010), who focuses specifically on the pedagogical applications of spoken corpora. Guilquin and Granger (2010) notably explore the synergy between Data-Driven Learning (DDL) and corpus methodologies. Complementing these works, Smith (2020) offers a comprehensive review of research concerning the use of corpora in foreign language education, especially within academic discourse contexts.

Despite the expanding literature base, a notable gap persists concerning the integration of corpus-based practices in teacher education. Specifically, there is limited insight into how pre-service teachers are trained to utilize corpora, and what forms of pedagogical intervention might effectively reshape their beliefs about the utility of CL in the classroom.

To address these issues, experiential learning emerges as a promising approach. As Gavioli (2009) suggests, learner motivation significantly increases when students are actively engaged in the creation and analysis of corpora. It is plausible to hypothesize that the same applies to future language educators. Accordingly, this study poses two interrelated research questions: (1) Does direct engagement with corpora and Natural Language Processing (NLP) tools influence teacher trainees' perceptions of corpus linguistics? (2) Does such engagement enhance their willingness to incorporate corpus-based methodologies in their future instructional practices?

2. Methodology

As Walsh (2023, p. 2) notes, "the main advantage of using corpora in teachers' education is that they provide an excellent point of access to complex phenomena related to teaching and learning." In alignment with this perspective, the present study engaged prospective language teachers in the compilation of corpora derived from informative texts. The objective was to identify the essential parameters for corpus construction and to explore how such corpora could be effectively employed in the instruction of English or Italian, using authentic language samples.

At the beginning of the experiment, university students possess held only a theoretical understanding of the pedagogical potential of corpora. However, through the process of constructing their own corpora and developing associated instructional strategies and classroom activities, their perceptions began to shift. This experiential learning approach aimed to foster a more nuanced and practical appreciation of corpus linguistics in language teaching. To measure attitudinal changes, two brief questionnaires were administered: one prior to the intervention and another upon its completion, both designed to elicit participants' views on the educational utility of corpora.

2.1. Participants

This study was inspired by the participants in the course named Corpus Linguistics Application in Teaching a Foreign Language, offered at the Faculty of Philosophy, University of Athens. The university students attending the course were from three different departments, namely the Department of English Language and Literature, the Department of Italian Language and Literature, and the Department of Greek Philology. To align with the study's objective of assessing the applicability of DDL strategies among future foreign language teachers, only those students who were concurrently undertaking teacher training were included in the research sample. Accordingly, the participant pool was limited to 4 second-year students from the English Department and 28 first-year students from the Italian Department based on their availability.

The participants were not pre-selected for the study; rather, they were drawn from a mixed-enrollment course, with inclusion criteria based solely on their engagement in foreign language education. While the students' target languages (English or Italian) were acknowledged, this variable did not play a decisive role in the study's design or analysis. All participants were required to complete individual assignments involving the creation of a native-speaker corpus (NS Corpus) and its pedagogical application in the foreign language classroom. These tasks were conducted using corpus analysis and Natural Language Processing (NLP) tools. The participants shared a relatively homogeneous educational background: all were Greek-speaking, of similar age, and enrolled at the same academic level.

2.2. Corpora

The corpora developed by the participants were composed of comprehension texts selected from the Panhellenic national examinations administered over the past decade (2013–2022). Although each student was tasked with constructing their own corpus, the final outputs were standardized by language. The English corpus comprised 3,572 words, while the Italian corpus consisted of 3,517 words. This uniformity facilitated comparative analysis and ensured assignment homogeneity. Despite the linguistic differences, the corpora were equivalent in size and genre, informative texts, thus providing comparable instructional potential across both language groups.

The Panhellenic examination texts used for corpus compilation are consistent in format across all languages, typically comprising informative texts of similar length—approximately 300 to 350 words—subject to minor annual variation. This consistency ensured that the corpora constructed by participants, despite being in different languages, remained comparable in both genre and scale.

In constructing their corpora, participants adhered to a standardized procedure. Each was instructed to assemble a mini corpus by retrieving the relevant reading comprehension texts, converting them into .txt format, and organizing them within a designated directory (Figure 1). Subsequently, they created a metadata file using an Excel spreadsheet, where they documented key attributes of each text, including the title, source URL, author or journalist's name, and publication date (Figure 2). In addition, they recorded quantitative metadata such as word count, sentence count, and other textual metrics to support corpus analysis.

Table 1.
The Features of Two Corpora

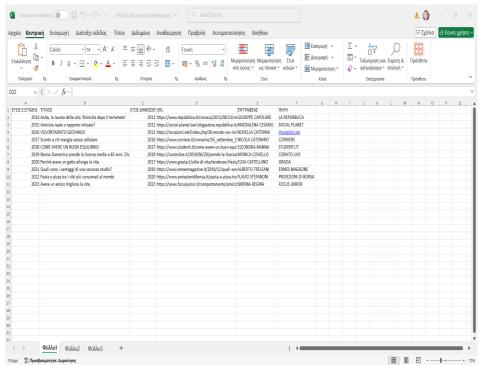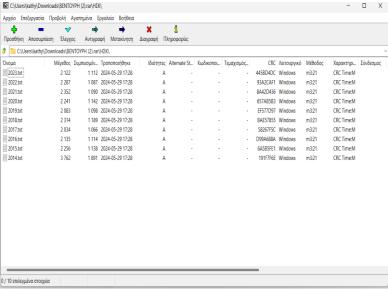| CORPORA	Corpus_it	Corpus_en
No of Words	3517	3572
Examined Texts	2013-2022	2013-2022
Corpora	Corpus_it	Corpus_en
No of Words	3517	3572
Examination Texts	2013-2022	2013-2022
Corpora	Corpus_it	Corpus_en
No of Words	3517	3572
Examination Texts	2013-2022	2013-2022

Figure 1. Corpus file

Figure 2. Corpus metadata

2.3. Educational proposals

The participants developed their assignments over a period of approximately four to five weeks. During the initial stages, particularly in the first week, there was sustained interaction with the instructor, as students sought guidance on pedagogical applications of corpora and expressed uncertainty regarding the use of NLP tools. To facilitate the process, two user-friendly and widely accessible tools, Voyant (Sinclair & Rockwell, 2016) and AntConc (Anthony, 2005), were recommended.

Upon completion, each student presented their assignment, which included both the constructed corpus and its instructional application. The educational themes addressed by the participants in their teaching proposals could be classified into the following categories:

Table 2.

Functions of NLP tools used by future teachers

NLP tool	Concordances	Collocates	Frequency lists	Other	TOTAL
Voyant	13	3	9	1	26
AntConc	4	1	1		6
TOTAL	17	4	10	1	32

Approximately half of the participants used their corpora primarily as databases for extracting linguistic data and designing instructional activities. The remaining participants employed corpus processing tools to develop interactive, hands-on learning experiences for their own students. The complexity of the educational proposals varied considerably.

Some assignments featured straightforward tasks, such as inferring the meaning of a word through the analysis of concordance lines. For example, one participant instructed students to examine concordances generated by Voyant to deduce the contextual meaning of the word *"respective"*. In a similar vein, an Italian-language activity involved guiding students to formulate a grammatical rule based on the usage patterns of the word *"perché"* within the corpus.

Other proposals encouraged learners to identify the most frequent lexical items in each text and use this information to classify the texts into thematic categories, such as youth, environment, or education. A further group of participants leveraged high-frequency vocabulary as a springboard for classroom discussion, prompting students to explore and engage with topics emerging organically from the corpus data.

2.4. Questionnaires

Following three introductory sessions—each lasting three hours—on the fundamentals of corpus linguistics, including key terminology and a historical overview of its applications, participants were administered an initial questionnaire. This instrument employed a three-point Likert scale ("yes," "maybe," "no"), which was deemed appropriate for a small-scale study of this nature (Taherdoost, 2019). The questionnaire design was informed by the work of Yoon and Hirvela (2004) but adapted to suit the specific pedagogical context of the present study. Notably, no demographic or personal background data were collected. The first questionnaire comprised four items

and was distributed prior to participants' engagement with corpus construction, NLP tool usage, and instructional design. These items assessed participants' initial perceptions of corpus linguistics, specifically in relation to its utility as a database for material extraction, its role in the foreign language teaching process, and their intent to adopt corpus-based methods in their future practice. The fourth item gauged perceptions of the effectiveness of combining NLP tools with small corpora for instructional purposes. To evaluate the internal consistency of the instrument, a reliability analysis was conducted using the Cronbach's alpha calculator from Cogn.IQ.org, yielding a coefficient of 0.9938, indicative of excellent reliability.

3. Results

Initial responses, as reflected in Table 3, revealed considerable skepticism toward corpus-based methods. Following an additional lesson focused on the use of NLP tools, participants were presented with a supplementary question regarding the perceived efficacy of integrating these tools with small-scale corpora for foreign language instruction. Despite this extended exposure, many participants expressed reservations, particularly in relation to unfamiliar corpora and the perceived complexity of the tools involved.

Table 3.
Answers to the First Questionnaire

Question	YES	MAYBE	NO
Are corpora useful as data bases for teaching material?	2	14	16
Are Corpora useful as tools for the teaching process?	2	13	17
Are you going to use Corpora as a FL teacher?	1	11	20
Are corpora compiled with NLP tools an effective method in FL teaching?	1	9	22

At the conclusion of the course, after all participants had completed their assignments and participated in both presenting their work and attending peer presentations, the same four questionnaire items were re-administered using the original three-point Likert scale. The responses, as detailed in the results, strongly indicate a marked shift in participants' perceptions regarding the pedagogical value of corpora and the practical utility of NLP tools in language instruction.

Table 4.
Answers to the Second Questionnaire

Question	YES	MAYBE	NO
Are corpora useful as data bases for teaching material?	26	2	4
Are Corpora useful as tools for the teaching process?	28	0	4
Are you going to use Corpora as a FL teacher?	27	1	4
Are corpora compiled with NLP tools an effective method in FL teaching?	25	3	4

To evaluate the significance of changes in participants' responses between the pre- and post-intervention questionnaires, a chi-square test of independence was employed. This statistical test determines whether the distribution of responses— "Yes," "Maybe," and "No"—differs significantly across the two datasets for each questionnaire item. The results of the chi-square analysis are presented below and provide empirical evidence of a statistically significant shift in participants' perceptions.

Table 5.
Chi-square Test

Chi-square statistic	164.67
Degrees of freedom (dof)	14
p-value	8.16×10^{-28} (essentially 0)

The extremely low p-value indicates a statistically significant difference in response patterns between the two datasets. Specifically, participants in the post-intervention questionnaire demonstrated a markedly higher inclination toward "Yes" responses across all items, suggesting a substantial positive shift in their attitudes toward the use of corpora and NLP tools in language teaching.

4. Discussion

The findings of this study corroborate existing research that underscores the pedagogical value of corpus-based tools, such as Voyant, in fostering data-driven language learning and critical thinking (Boulton, 2010; Römer, 2006). Consistent with previous literature (Stubbs, 2004; Leech, 2014; Wulf & Baker, 2021), participants in this study utilized concordance lines and word frequency data to promote learner autonomy and inductive reasoning. Notably, the diversity of instructional applications observed, ranging from lexical inference to thematic classification and discourse-based activities, suggests a broader pedagogical engagement than is commonly reported in earlier studies.

A particularly significant development was the evident transformation in participants' attitudes toward corpus tools over the course of the study. Initially approached with skepticism, these digital resources gradually gained acceptance as participants engaged directly with the tools. This hands-on experience appeared to demystify their operation and relevance, leading participants to reframe these tools not as overly technical, but as pedagogically useful. This shift aligns with prior findings indicating that practical application is instrumental in building teacher confidence with digital technologies (Yoon & Hirvela, 2004).

Despite these positive outcomes, certain limitations were observed. While the corpus-based tasks were generally well-received, few participants demonstrated critical pedagogical reflection beyond

surface-level tool usage. Most focused on technical functionality or immediate classroom applicability, rather than engaging with deeper educational dimensions such as learning outcomes, student agency, or long-term curriculum integration. This observation echoes concerns in the literature about the gap between technological implementation and pedagogical theorization (Bennett & Oliver, 2011).

Moreover, although practical exposure helped mitigate initial resistance, it alone is insufficient to catalyze sustained pedagogical transformation. Without structured opportunities for theoretical engagement and reflective practice, there is a risk that corpus tools may be adopted in a merely instrumental manner. Future instructional designs should therefore embed critical reflection and pedagogical theory to ensure that corpus-based approaches are integrated meaningfully and sustainably into language teaching practices.

5. Conclusion

The results of the second questionnaire, supported by the chi-square analysis, provide strong evidence that corpus-based instruction is perceived as a valuable pedagogical tool by the participating future teachers. All 32 participants successfully completed the assigned tasks, including the construction of a corpus, the creation of a detailed metadata spreadsheet, and the application of NLP tools to design foreign language teaching activities. This comprehensive engagement highlights a key finding: when teacher trainees are directly involved in the creation and application of corpora, their motivation to utilize such tools in pedagogical contexts increases significantly—a point previously emphasized by Guilquin and Granger (2010). The hands-on experience contributed to a notable reduction in initial skepticism toward both corpora and NLP tools, affirming the efficacy of experiential learning in overcoming resistance to new instructional methodologies.

These findings support the integration of Corpus Linguistics into teacher training programs as a means of cultivating a new generation of language educators who are receptive to data-driven instructional practices. The observed shift in participants' attitudes underscores the transformative potential of corpus-based approaches in making language education more dynamic, reflective, and effective.

However, this study is not without limitations. It involved a relatively small and homogeneous group of participants, and their corpora were limited in size. Nonetheless, small, specialized corpora offer unique advantages, particularly in their ability to address specific pedagogical questions and maintain a close link between language use and context (Reppen, 2010; Ross, 2018; Vaughan & Clancy, 2013).

Given the limited scope of the sample, caution must be exercised in generalizing the findings. Future studies should aim to include a larger and more diverse cohort of participants to ensure the robustness of results. Moreover, longitudinal research is necessary to assess the sustained impact of corpus-based training on instructional practices and teacher beliefs. Understanding how to systematically foster an appreciation for authenticity and learner autonomy through corpus methodologies remains a key challenge for teacher educators.

References

1. Anthony, L. (2005). AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. *IPCC 2005. Proceedings. International Professional Communication Conference.* Limerick, Ireland, (pp. 729-737). DOI: https://doi.org/10.1109/IPCC.2005.1494244

2. Aston, G. (2002). The learner as corpus designer. In B. Kettemann, & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 9-25). Amsterdam: Rodopi.

http://www.sslmit.unibo.it/~guy/graz.htm

3. Bennett, S., & Oliver, M. (2011). Talking back to theory: The missed opportunities in learning technology research. *Research in learning Technology*, *19*(3), 179-189. https://doi.org/10.3402/rlt.v19i3.17108

4. Boulton, A., & Tyne, H. (2014). Corpus-based study of language and teacher education. *The routledge handbook of educational linguistics.* Routledge.

5. Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning, 60*(3), 534-572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

6. Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. ReCALL, 21(1), 37-54.

7. Boulton, A. (2008). Bringing corpora to the masses: Free and easy tools for language learning. In N. Kübler (Ed.), *Corpora, language, teaching, and resources: From theory to practice* (pp. 1-19). Peter Lang. http://hal.archives-ouvertes.fr/docs/00/32/69/80/PDF/XXXX_boulton_TaLC_interdisciplinary.pdf

8. De Cock, S. (2010). Spoken learner corpora and EFL teaching. *Corpus-based approaches to English language teaching*, (pp. 123-137).

9. Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language awareness*, *17*(1), 25-43. DOI: https://doi.org/10.2167/la414.0

10. Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for specific purposes*, 24(3), 321-332.

11. Gavioli, L. (2009). Corpus analysis and the achievement of learner autonomy in interaction. In L. Lombardo (Ed.), *Using corpora to learn about language and discourse* (pp. 39-71). Peter Lang.

12. Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? *The Routledge handbook of corpus linguistics*. Routledge.

13. Karlsen, P. H. (2021). *Teaching and Learning English through Corpus-based approaches in Norwegian Secondary Schools: identifying obstacles and a way forward*. [Doctoral dissertation]. https://brage.inn.no/innxmlui/handle/11250/2829366

14. Leech, G. (2014). Teaching and language corpora: A convergence. *Teaching and language corpora* (pp. 1-24). Routledge.

15. Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

16. Römer, U. (2006). *Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. Zeitschrift für Anglistik und Amerikanistik, 54*(2), 121-134.

17. Ross, D. (2018). Small corpora and low-frequency phenomena: Try and beyond contemporary, standard English.*Corpus*, (18). DOI: https://doi.org/10.4000/corpus.3574

18. Sinclair, S., & Rockwell, G. (2016). *Voyant Tools.*

19. Smith, S. (2008). DIY corpora for vocabulary learning. 第二辑, 3-26. https://xuebao.zyufl.edu.cn/_upload/article/files/6a/a9/a93a08fc4191b0e67f3b5051e874/64381a39-52a6-4776-ace7-0f31c242b341.pdf

20. Vaughan, E., & Clancy, B. (2013). Small corpora and pragmatics. In *Yearbook of corpus linguistics and pragmatics 2013: New domains and methodologies* (pp. 53-73). Dordrecht: Springer Netherlands.

21. Stubbs, M. (2004). Language corpora. In A. Davies, & C. Elder (Eds.), *The handbook of applied linguistics*, (pp. 106-132). Blackwell Publishing Ltd. DOI: https://doi.org/10.1002/9780470757000

22. Taherdoost, H. (2019). What is the best response scale for survey and questionnaire design? Review of different lengths of rating scale/attitude scale/Likert scale. *International Journal of Academic Research in Management*, *8*(1), 1-10. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3588604

23. Walsh, S. (2023). Using corpora for English language teacher education (ELTE). *Roczniki Humanistyczne, 71*(10S), 193-212. https://doi.org/10.18290/rh237110sp-10

24. Wulff, S., & Baker, P. (2021). Analyzing concordances. *A practical handbook of corpus linguistics*. Springer International Publishing.

25. Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2. *Journal of second language writing*, *13*(4), 257-283. https://doi.org/10.1016/j.jslw.2004.06.002